Multivariate Analysis of a Data Matrix Containing A-DNA and B-DNA Dinucleoside Monophosphate Steps: Multidimensional Ramachandran Plots for Nucleic Acids

M. L. M. BECKERS, L. M. C. BUYDENS

Laboratory for Analytical Chemistry, Faculty of Science, Catholic University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Received 18 February 1997; accepted 30 November 1997

ABSTRACT: A method to construct the equivalent of multidimensional Ramachandran plots for nucleic acids on the basis of singular value decomposition (SVD) is presented. For this purpose, a data matrix containing 244 DNA dinucleoside monophosphate steps, represented by nine torsion angles, was decomposed into a score and loading matrix. It is shown that biplots, containing both score points and loading vectors, provide a simple tool to interpret the principles of DNA class separation. Scores separate the data matrix into one A-DNA class, two different B-DNA classes, and one so-called crankshaft class. Loading vectors correlate torsion angles. The projections of scores on loading vectors indicate which torsion angles play a dominant role in DNA class separation. The results of the biplots are supported by (simple) physical interpretations. From a three-dimensional score space the nine original torsion angles can be reconstructed. Hence, the potential to create the multidimensional equivalent of a Ramachandran plot is available; that is, forbidden and accessible regions in the reduced space reflect these same regions in the nine-dimensional original space. © 1998 John Wiley & Sons, Inc. J Comput Chem 19: 695–715, 1998

Keywords: nucleic acid; multivariate analysis; Ramachandran plots

Correspondence to: L. M. C. Buydens

CCC 0192-8651 / 98 / 070695-21

Introduction

The biological activity of biomacromolecules, such as proteins and nucleic acids, is largely defined by their three-dimensional (3D) spatial structure, or conformation. Conformational analysis requires a proper means of representing trial structures. In this study, the focus is on nucleic acids. Common representations for both single strand and double strand nucleic acids are atomic coordinates and torsion angles. Dickerson et al.[1] gave definitions and nomenclature for different structure parameters that describe both base pair structures and the structure of individual bases in a single strand.

Even for relatively small molecules the degrees of freedom in conformational analysis can be very large. For instance, a dinucleoside monophosphate step in torsion angle space can be represented by nine torsion angles which form the *dimension* of the search space. When each of the torsion angles is allowed to vary in a range of 60° and a search is performed with a step size of 6°, hence the *resolution* is 10, then the search space covers 10^9 conformations. More generally, the size of the search space is defined by the resolution to the power of the dimension. Obviously a reduction in the number of variables to optimize, hence a reduction in the dimension, would decrease the size of the search space in the conformational analysis. One way to reduce the dimension is to search for relations between the variables.

Most biomacromolecules occur in certain preferred conformations. Nucleic acid double helices can be of, for example, A-DNA or B-DNA helix types, and proteins are often comprised of substructures such as β-sheets, etc. Therefore, obviously, there are strong underlying relations between parameters that define a conformation. Hence, besides the reduction of the search space in conformational analysis the study of relations between torsion angles (and/or other structure parameters) may reveal interesting information about the structural behavior of biomacromolecules. Attempts have been made to calculate correlation coefficients between different structural (helical) parameters for nucleic acids.[2,3] In this article, we focus on the correlation between torsion angles in nucleic acids. Investigations of the relation between the two main torsion angles that define a protein backbone have led to the famous Ramachandran plots.[4]

Because the backbone of nucleic acids is defined by more than two torsion angles the construction of the equivalent of Ramachandran plots is more complicated. However, recently, Mooren et al. calculated accessible and forbidden areas for torsion angle combinations on the basis of steric hindrance effects. Pairwise interactions between torsion angles are investigated while keeping zero, one, and two torsion angles fixed, respectively[5] (see Appendix). Early studies on the relationships between torsion angles in nucleic acids can be found in refs.[6–9] Altona and Sundaralingam introduced the now widely used principle of pseudorotation. They found that the relation between torsion angles ν_0 to ν_4 of the furanose ring could be modeled by using two parameters, namely the pucker phase angle P and the pucker amplitude ϕ.[10] Fratini et al. synthesized four variants of a B-DNA dodecamer and determined their crystal structure. Pairwise correlation coefficients between torsion angles of each dodecamer were calculated. The torsion angle pairs that had a correlation coefficient larger than 0.5 are summarized as (see Fig. 1 for the torsion angles used)[11]:

Fratini et al.	$\chi-\delta$	$\chi-\zeta$	$\delta-\zeta$	$\epsilon-\zeta$	$\epsilon-\beta$	$\zeta-\beta$
Conner et al.	$\chi-\delta$	$\chi-\epsilon$	$\chi-\alpha$	$\epsilon-\alpha$	$\zeta-\alpha$	$\zeta-\gamma$
	$\alpha-\beta$	$\alpha-\gamma$				

Conner et al. synthesized an A-DNA tetramer and determined its crystal structure and the corresponding torsion angles. They then collected the torsion angles of four A-DNA structures and five B-DNA structures with known crystal structures. Torsion angle pairs with a correlation coefficient larger than 0.5 for the five B-DNA structures are the same as those reported by Fratini et al. For the five A-DNA structures, correlation coefficients larger than 0.5 were reported by Conner et al.[12] They are given in the previous table.

Both Fratini et al. and Conner et al. calculated correlation coefficients for a rather small amount of data points derived from structures with similar conformations. Besides, most of the earlier studies focused on linear regression between pairs of torsion angles. A recent study by Schneider et al.[13] extensively visualized pairwise correlations between torsion angles of structures collected from the Nucleic Acid Databank (NDB).[16]

However, finding relations between torsion angles of nucleic acids is clearly a multivariate prob-

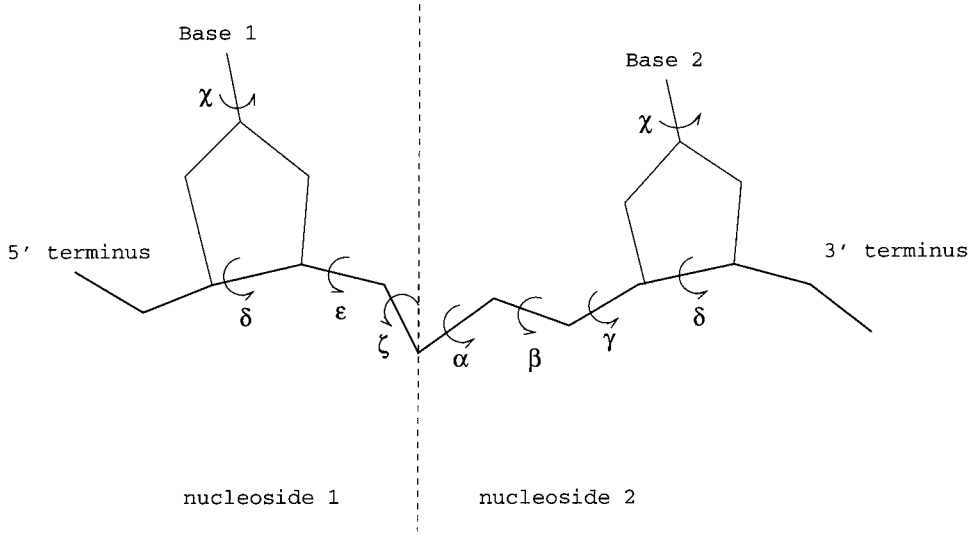

FIGURE 1. Torsion angle representation of the dinucleoside monophosphate step used in this study. The backbone conformation is represented by torsion angles α, β, γ, ϵ, and ζ. The complete conformation of the furanose ring is not taken into account. Instead, it is represented by δ. The orientation of the base with respect to the sugar ring is given by torsion angle χ.

lem. Consider again the dinucleoside monophosphate (DM) step, which is represented by nine torsion angles. Trying to find relations between torsion angles then comes down to collecting a set of m DMs. When the torsion angles of each DM are known, the data set can be represented as a matrix of m objects described by n variables. Possibly all, or some, of the variables influence each other and are not only pairwise correlated but have multiple correlations. Besides, when well-defined pairwise relations are known it is of interest to investigate how they might be interrelated. Hence, it is worthwhile to use methods capable of analyzing all variables at the same time. Pearlman and Kim made a first attempt in this manner.[14] They collected torsion angles of four A-DNA structures, two B-DNA structures, four Z-DNA structures, and a tRNA molecule. Empirical multiple correlation functions for three torsion angles in the nucleic acid backbone have been reported: $\alpha_i = f(\zeta_{i-1}, \beta_i, \gamma_i, \epsilon_i)$; $\delta_i = f(\epsilon_i, \zeta_i, \chi_i, \beta_{i+1})$; and $\epsilon_i = f(\zeta_i, \chi_i, \beta_{i+1})$. The relations, which include linear and nonlinear (sin and cos) terms, were found by means of an exhaustive multiple regression analysis on the data matrix.

In this study, we present a multivariate investigation into the relations between torsion angles in nucleic acids. A data matrix with 244 objects was constructed. Each object is a DM represented by a vector of nine torsion angle values. DMs were collected from structures with known crystal conformations for both A-DNA and B-DNA. The B-DNA DMs are subdivided in a B_I-DNA family of rotamers and a B_{II}-DNA family of rotamers according to the definition by Privé et al.[15] The data matrix was decomposed into a score matrix and a loading matrix via singular value decomposition (SVD). Good DNA class separation is obtained in the score plots, whereas relations between variables are estimated from loading plots. We show that biplots, which depict both scores and loadings in the same plot, not only provide class separation and the relation between variables but also give insight into how class separation is caused by the variables. The fact that clusters of DM classes are formed in scores plots suggests that the spaces between the clusters can be considered forbidden areas. Therefore, a procedure to create an equivalent of multidimensional Ramachandran plots for nucleic acids on the basis of score plots is suggested.

The most important goal of the research described in this article is to explore the possibilities of an easy-to-use multivariate technique, provided by SVD, in elucidating classes of objects and the corresponding relation between variables in a DM data matrix. Therefore, a well-defined data set is derived from molecules whose structure has been published. It is demonstrated that the method works well in describing the data set and may therefore be used in the structural analysis of more exotic molecules. By relating the SVD results to the

interpretation of structural parameters, such as base stacking and steric hindrance, the basis for the equivalent of a multidimensional Ramachandran plot is provided.

Materials and Methods

DATA SET

Table I summarizes the data used in this study It is based on the study of conformational mobility of DM steps by El Hassan and Calladine.[3] All references in which the crystal structures of sequences given in Table I were published are given in ref. 3. The reader is also referred to this reference for information on the space group that was used in the crystallographic studies, the resolution (between 1.4 and 2.6 Å) and the *R*-factor (between 11.5% and 23.2%). Most of the structures used in this study are available from the NDB (see ref. 16).*

Each single-strand sequence was subdivided in DM steps. We describe the DM steps by the nine torsion angles depicted in Figure 1. Hence, the data matrix contains nine columns or variables. The individual DMs, represented by nine variables, are the rows or objects, of the data matrix. DMs containing the bases inosine and uracil as well as DMs with bases in mismatched base pairs were not added to the data matrix (they are indicated in boldface in Table I). The reason for this is that we initially wanted to create a data matrix with ''common'' DNA characteristics. Much information about the common features of DNA is available from the literature. Hence, we have a proper means to test whether our method works well. Thereafter, it may be used for other molecules, such as RNA. Also, a small number of DMs that had one or more torsion angle combination(s) in ''forbidden areas,'' as described elsewhere[5] (see Appendix, which summarizes the method used in ref. 5), were not added to the data matrix. These entries appear to have torsion angle combinations that theoretically cause (large) van der Waals overlap for some atom pairs. We analyzed the distribution of torsion angles in advance. From this, it became clear that intermediate torsion angle com-

TABLE I.
Sequence of Structures from which DM Steps Were Taken and Corresponding Helix Types.[1]

Sequence	Helix type	NDB code
Dodecamers		
d(CCGTACGTACGG)	A	adl045
d(GCGTACGTACGC)	A	adl046
d(CGC**I**AATT**A**GCG)	B	bdlb10
d(CGCGAATTCGCG)	B	DODBAJ[a]
(CGCAAATTTGCG)	B	bdl038
d(CGCGAATTCGCG)	B	bdbl04
d(CG**A**AATT**C**GCG)	B	DODBAC[b]
d(CGC**G**AATT**T**GCG)	B	bdl009
d(CGC**A**AATT**I**GCG)	B	bdbl41
d(CGC**G**AATT**G**GCG)	B	bdl046
Decamers		
d(CGATCGATCG)	B	bdjb48
d(CGGTATACGC)	A	ahj040
d(GCGTATACGC)	A	ahj043
d(GCGTATACGC)	A	ahj044
d(CGATCGATCG)	B	bdj025
d(CCAGGCCTGG)	B	bdj017
d(CCAGGCCTGG)	B	bdjb27
d(CCAA**CI**TTGG)	B	bdjb43
d(CCAA**CI**TTGG)	B	bdjb44
d(CCAA**GA**TTGG)	B	bdj008
d(CCAACGTTGG)	B	bdj019
d(CGATTAATCG)	B	bdj031
d(CGATATATCG)	B	bdj036
Octamers		
d(GG**I**GC**T**CC)	A	adhb17
d(GGGTACCC)	A	OCTAAR[c]
d(CCCCGGGG)	A	OCTAAA[d]
d(GCCCGGGC)	A	adh008
d(GCCCGGGC)	A	adhp36
d(GG**G**GC**T**CC)	A	OCTAAJ[e]
d(CTCTAGAG)	A	adh020
d(GG**UAUA**CC)	A	adhb11
d(GGG**GT**CCC)	A	OCTAAK[f]
d(GGGATCCC)	A	OCTAAG[g]
d(GGGGCCCC)	A	OCTAAI[h]
d(GTACGTAC)	A	adh024
Tetramer		
d(CCGG)	A	addb01

[1] DM steps from mismatched base pairs and base pairs containing inosine and uracil are indicated in boldface.

* Torsion angles for structures encoded by DODBA* and OCTAA* can be found in: (a) H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and D. E. Dickerson, *Proc. Natl. Acad. Sci. USA*, **78**, 2179 (1981); (b) W. N. Hunter, T. Brown, and O. Kennard, *Nucl. Acids Res.*, **15**, 6589 (1987); (c) M. Eisenstein, F. Frolow, Z. Shakked, and D. Rabinovitch, *Nucl. Acids Res.*, **18**, 3185 (1990); (d) T. E. Haran, Z. Shakked, A. H.-J. Wang, and A. Rich, *J. Biomol. Struct. Dynam.*, **5**, 199 (1987); (e) W. N. Hunter, G. Kneale, T. Brown, D. Rabinovitch, and O. Kennard, *J. Mol. Biol.*, **190**, 605 (1986); (f) G. Kneale, T. Brown, O. Kennard, and D. Rabinovitch, *J. Mol. Biol.*, **186**, 805 (1985); (g) H. Lauble, R. Frank, H. Bloecker, and U. Heinnemann, *Nucl. Acids Res.*, **16**, 7799 (1988); (h) M. McCall, T. Brown, and O. Kennard, *J. Mol. Biol.*, **183**, 385 (1985).

binations are also present. The only criterion for leaving structures out is the fact that they will lead to unavoidable van der Waals overlap according to the Mooren criterion. This is, for instance, the case for very low values of ϵ (see Appendix). Because we do not know the reason for these rare ϵ values (low resolution of the X-ray crystallography measurements perhaps?) the few structures in which such a low ϵ was present were left out. This does not mean that combinations with intermediate ϵ and ζ values (between B_I-DNA and B_{II}-DNA) were left out. The same holds for combinations with intermediate δ and χ values (between A-DNA and B-DNA). This resulted in a total of 244 DM entries. The first 144 entries are B-DNA entries. The B-DNA entries are subdivided in a B_I-family of rotamers with $\epsilon(tr)/\zeta(g-)$ and a B_{II}-family of rotamers with $\epsilon(g-)/\zeta(tr)$.[†] The remaining 100 entries represent A-DNA conformations.

DATA PRE-TREATMENT

When collecting torsion angles we found that some investigators reported torsion angles according to a 0–360° range, whereas others reported according to a −180–180° range. To make our data matrix consistent we had to choose a range. Because the former range worked slightly better for our scaling procedure we chose to work with that range.

When performing multivariate techniques the data are usually scaled. Mean-centering and autoscaling are among the most often used scaling techniques. Several scaling techniques were attempted in this study. Because some of the variables have a non-normal distribution it was decided to scale the data in each column by subtracting its corresponding median.

CORRELATION MATRIX

The coherence between two sets of measurements can be expressed by the correlation coefficient. Pairwise calculation of correlation coefficients for n variables in a data matrix results in an $n \times n$ symmetric correlation matrix. In this study, correlation coefficients are used in validating the results obtained with the singular value decomposition method described next.

[†] The expressions in parentheses are used to indicate the corresponding torsion angle ranges: g^+: 60 ± 60°; tr: 180 ± 60°; g^-: 300 ± 60°.

SINGULAR VALUE DECOMPOSITION

In singular value decomposition (SVD), the original $m \times n$ data matrix, $\mathbf{X}$, is decomposed into an $m \times n$ matrix of row-singular vectors, $\mathbf{U}$; an $n \times n$ diagonal matrix of associated singular values (square roots of the eigenvalues), Λ; and an $n \times n$ matrix of column-singular vectors, $\mathbf{V}$. $\mathbf{U}$ contains eigenvectors of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{V}$ contains eigenvectors of $\mathbf{X}^T\mathbf{X}$ (in the economy-sized decomposition only n columns of the $m \times m$ matrix $\mathbf{X}\mathbf{X}^T$ are involved).

From the point of view of a geometrical interpretation, the kth column of $\mathbf{U}$ defines the kth singular vector in row-space S^m. Similarly, the kth column of $\mathbf{V}$ defines the kth singular vector in column-space S^n. The singular vectors in both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. Therefore, they define an orthogonal system of basis vectors in each of the dual spaces S^m and S^n. The row-space S^m can be defined as the coordinate space in which the n columns of the $m \times n$ matrix $\mathbf{X}$ can be represented as a pattern of n points. Similarly, column-space S^n is the coordinate space in which the m rows of $\mathbf{X}$ can be represented as a pattern of m points. For an extensive description of singular value decomposition and its geometrical interpretation the reader is referred to ref. 17:

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T \tag{1}$$

Coordinates for the m rows of $\mathbf{X}$ in the reduced row-space, or scores, are given by:

$$\mathbf{S} = \mathbf{U}\Lambda^{\alpha} \tag{2}$$

Coordinates of the n columns of $\mathbf{X}$ in the reduced column space, or loadings, result from:

$$\mathbf{L} = \mathbf{V}\Lambda^{\beta} \tag{3}$$

where α and β are 0, 0.5, or 1, respectively. Sometimes $\mathbf{U}$ and $\mathbf{V}$ are referred to as principal components (PCs). Principal component analysis (PCA) is commonly used to elucidate the structure in a data matrix $\mathbf{X}$ and amounts to $\alpha = 1$. PCA tries to represent the most important aspects of the original variables by means of a smaller number of newly created variables. The original $m \times n$ data matrix, $\mathbf{X}$, is decomposed into an $m \times n$ scores matrix, $\mathbf{S}$, and an $n \times n$ loadings matrix, $\mathbf{L}$:

$$\mathbf{X} = \mathbf{S}\mathbf{L}^T \tag{4}$$

When the first PC has been defined, the second PC is chosen to be orthogonal to the first PC, etc. The total amount of variation, explained by a PC, is

called the ''eigenvalue,'' λ^2, of the PC. The first PC explains the most variance and the last PCs usually explain very little variance. Instead of interpreting score and loading plots separately, which is usually done, it is of interest to look at a multidimensional representation of the data, especially in the case of DMs. For this purpose, so-called biplots, in which scores and loadings are depicted in the same plot, can be used.[19,20]

For $\alpha = 0$ and $\beta = 1$ a so-called column preservation biplot is created (in ref. 18 the investigators call these ''GH biplots''):

$$\begin{aligned} \mathbf{S} &= \mathbf{U} \\ \mathbf{L} &= \mathbf{V}\mathbf{\Lambda} \end{aligned} \tag{5}$$

Distances between the representations of columns are preserved. The distances, d_j, from the origin to a loading, $\mathbf{l}_j$, hence the length of the vector, is a measure of information content of the corresponding column; that is, vectors with greater lengths contribute more to a PC than vectors with short lengths. Not only distances from the origin to a loading but also the distance between two loadings can be calculated. This can be combined to derive the angle between two loading vectors. The cosine of the angle between vectors, $\mathbf{l}_j$ and $\mathbf{l}_{j'}$ approximates the correlation coefficient between columns j and j':

$$\begin{aligned} d_j &= \left(x_j^T x_j\right)^{\frac{1}{2}} \\ d_{j'} &= \left(x_{j'}^T x_{j'}\right)^{\frac{1}{2}} \\ d_{jj'} &= \left[(x_j - x_{j'})^T (x_j - x_{j'})\right]^{\frac{1}{2}} \\ \cos(\theta_{jj'}) &= \frac{d_{jj'}}{d_j d_{j'}} \end{aligned} \tag{6}$$

$\alpha = 1$ and $\beta = 0$ amounts to a *row preserving* (or in ref. 18 the JK) biplot. In this case, the distances between representations of the rows are preserved; that is, the scores can be interpreted.

S and **L** have the same weights when $\alpha = 0.5$ and $\beta = 0.5$ (SQ biplot in ref. 18):

$$\begin{aligned} \mathbf{S} &= \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}} \\ \mathbf{L} &= \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}} \end{aligned} \tag{7}$$

Eq. (4) shows that the original data can be reconstructed from a score and loading matrix. The necessary condition is that $\alpha + \beta = 1$. It can be proven that reconstruction is also possible by using perpendicular projections of the scores, $\mathbf{s}_i$, on the loadings, $\mathbf{l}_j$. Hence, biplots can be used as point-vector plots. A high value of the perpendicular intersection of the projection of a point $\mathbf{s}_i$ on a vector $\mathbf{l}_j$, represents a high value for point i in column j.

When performing PCA, or SVD, one usually determines the true dimensionality, or pseudorank, of the data; that is, the number of independent variables in the data matrix, hence, the number of relevant PCs, or variables, that describe the data matrix. One could simply look at how much variance is explained by the respective PCs. It is common use to take successive PCs into account until about 95% of the variance is explained. Other approaches are based on the eigenvalues themselves. Graphically, SCREE plots, in which the size of the eigenvalue is plotted against its index,[21] are popular. Widely used are Malinowski's *F*-test and reduced eigenvalue (REV) criterion; see eq. (8)[22]:

$$REV_p = \frac{\lambda_p^2}{(m - p + 1)(n - p + 1)} \tag{8}$$

Here p is the pth eigenvalue, m is the number of objects, and n is the number of variables. The pseudorank is determined from a plot of the *REV* against its PC. It is the index of the eigenvalue after which the *REV* becomes constant.

SOFTWARE

Basic statistical methods and singular value decomposition, including the calculations necessary for constructing biplots, were performed with Matlab for Unix workstations (version 4.2c) by The MathWorks, Inc.

Results and Discussion

SINGULAR VALUE DECOMPOSITION

Pseudorank Estimation

The eigenvalues of the scaled data matrix were extracted from the SVD procedure (Table II).

From a plot of the *REV* values against the index of the eigenvalues the number of independent variables in the original data matrix is estimated to be four. The first four PCs explain about 93% of the variance, whereas PC#1 and PC#2 already account for 75% of the variance. Hence,

TABLE II.
Eigenvalues, Percentage Variance Explained, and *REV* after SVD on Scaled Data Matrix.

PC#	λ^2 ($\cdot 10^6$)	Variance explained (%)	Σ[Variance explained] (%)	*REV*
1	1.0656	57.88	57.88	485.2
2	0.3189	17.32	75.20	164.0
3	0.2211	12.01	87.21	130.5
4	0.1097	5.96	93.17	75.9
5	0.0526	2.86	96.03	43.9
6	0.0356	1.93	97.96	37.2
7	0.0174	0.95	99.01	24.4
8	0.0136	0.74	99.75	28.7
9	0.0066	0.35	100.00	27.8

the structure of the original data is well preserved in the first few PCs.

Scores

According to $\alpha = 1$ and $\beta = 0$, the eigenvectors of $\mathbf{X}\mathbf{X}^T$, corresponding to the scores, were obtained. Figure 2 depicts the score plots of the first three PCs. A clear separation of three classes of DMs is seen in Figure 2a. Clearly, PC#1 separates A-DNA from B-DNA. There also is a slight separation of B_I-DNA from B_{II}-DNA on PC#1. PC#2 separates B_I-DNA from B_{II}-DNA. Figure 2b shows that, on PC#3, a small fourth class of objects is distinguished. This small group of objects exhibits the so-called crankshaft effect. This means that the combination $\alpha(g-)/\gamma(g+)$ is switched to $\alpha(tr)/\gamma(tr)$. Although crankshaft effects are most often reported for A-DNA, Figure 2b also shows a B-DNA DM with a high PC#3 score. Score plots of PCs greater than three did not show any additional class separations. From this class separation it can be deduced that certain areas in the score space are ''forbidden'' or at least not favorable. Therefore, it should be worthwhile to transform these areas into the original variable space. It is demonstrated later in this study that this transformation is possible and hence score plots can be used as a multidimensional equivalent of the Ramachandran plot.

How are the scores on the PCs related to the original variables? One might compare the distributions of the scores with the distribution of the original variables in the data matrix. In Figure 3 the distribution of PC#1 is given. As can be seen in the figure, its behavior is visually comparable to torsion angle δ. It also is comparable to torsion angle χ. In the same manner it can be visualized that the distribution of ϵ is strongly represented by PC#2 and the distribution of γ by PC#3. The relation between the PCs and the original variables is more quantitatively expressed by their corresponding correlation coefficients in Table III.

The combination of the score plots with these correlation coefficients leads to the following preliminary conclusion:

PC	Separation	Torsion angles involved
1	A-DNA/B-DNA B_I-DNA/B_{II}-DNA	 $\chi\delta\zeta\gamma$
2	B_I-DNA/B_{II}-DNA	$\epsilon\beta$
3	crankshaft	$\alpha\gamma$

PC#3 shows only a strong correlation with α and γ. Taking the sign of the corresponding correlation coefficients into account reveals that a strong negative correlation exists between α and γ. Obviously, this relation separates the crankshaft entries from the others. On PC#2, B_I-DNA/B_{II}-DNA separation occurs. Table III indicates that ϵ and β have rather high correlations with PC#2, whereas there is also reasonable correlation of PC#2 with ζ and χ_2. More torsion angles are highly correlated with PC#1. To further specify correlations between variables, loading plots are examined.

Loadings

By using $\alpha = 0$ and $\beta = 1$, the eigenvectors of $\mathbf{X}^T\mathbf{X}$, corresponding to the loadings, were obtained. Figure 4 depicts the loading plots of the first three PCs.

The χs, δs, and ζ form the larger part of PC#1, and ϵ and ζ form the larger part of PC#2 as do α and γ for PC#3. This compares reasonably well to the results in Table III. A difference is seen on

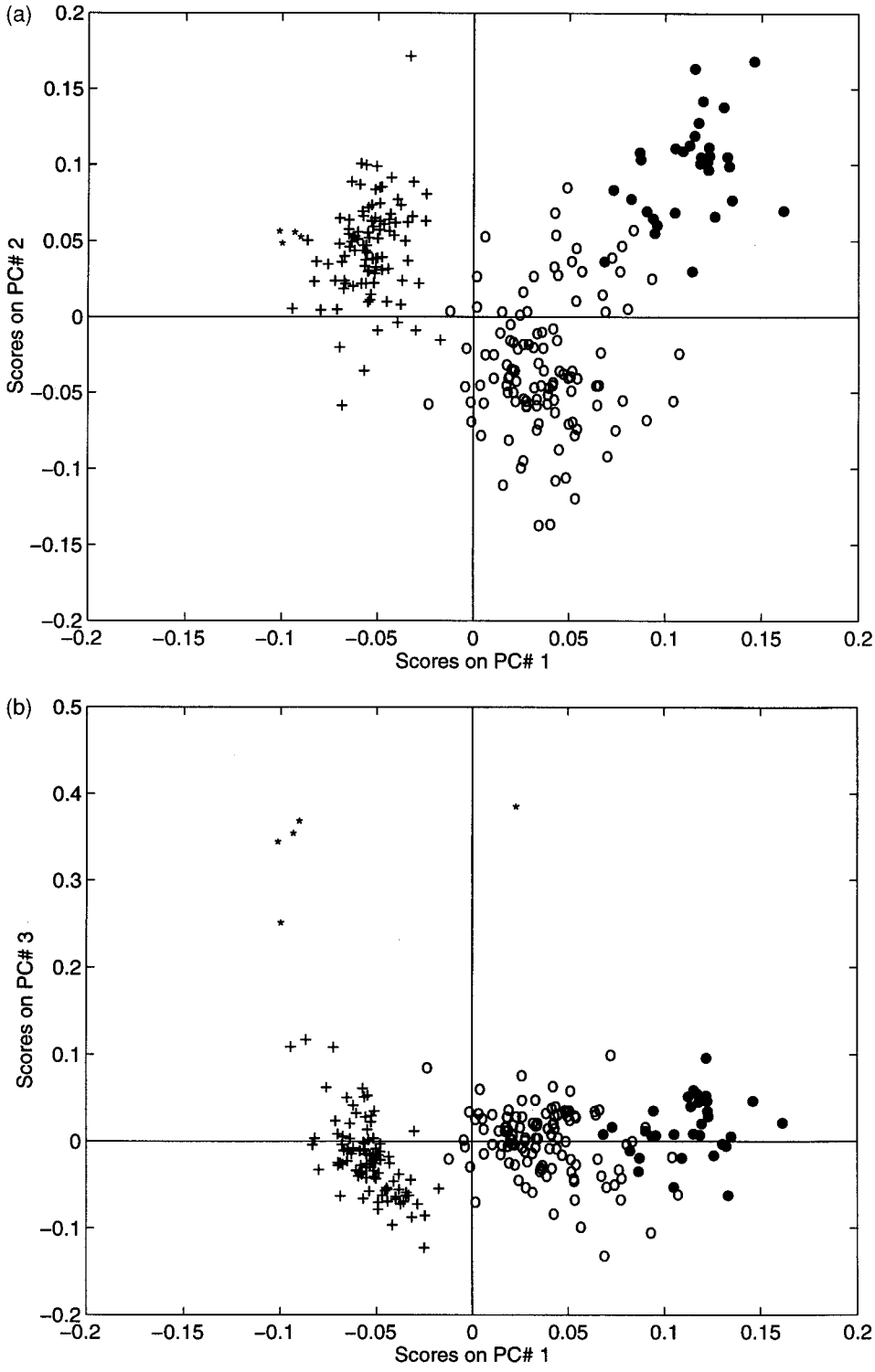

FIGURE 2. A-DNA (+); B_I-DNA (○); B_{II}-DNA (●); crankshafts (*). (a) Scores on PC#2 versus scores on PC#1; (b) Scores on PC#3 versus scores on PC#1.

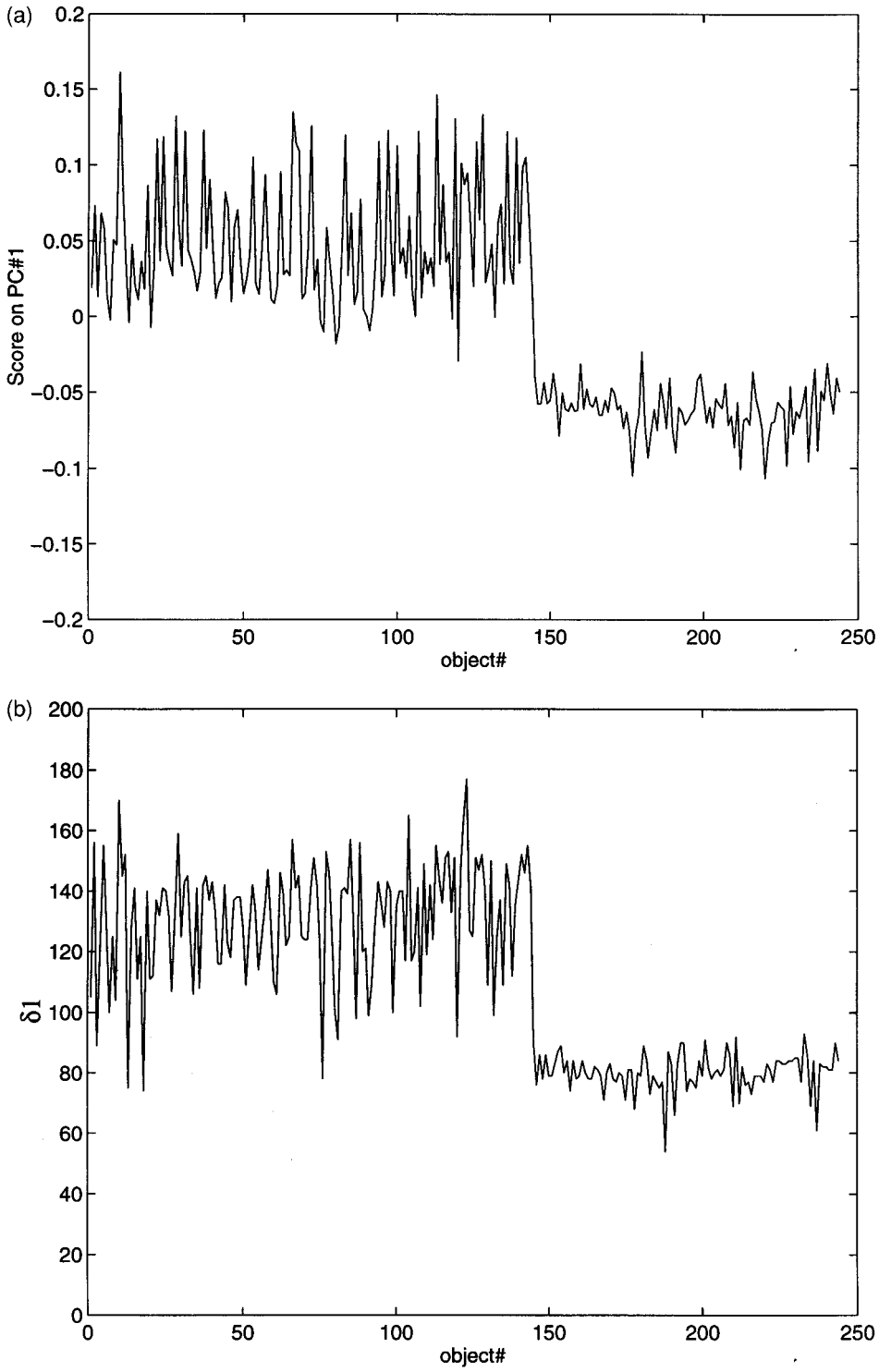

FIGURE 3. (a) Scores on PC#1 against data matrix entry. (b) Torsion angle δ against data matrix entry.

TABLE III.
Correlation Coefficients between the Original Variables and PC#1 to PC#5.

Variable	PC#1	PC#2	PC#3	PC#4	PC#5
χ_1	0.90	−0.24	0.20	−0.26	0.06
δ_1	0.89	−0.17	0.18	−0.30	0.12
ϵ	0.30	0.82	−0.03	0.33	0.33
ζ	−0.87	−0.46	−0.06	−0.01	−0.02
α	0.42	0.00	−0.88	−0.16	−0.16
β	−0.48	−0.56	0.12	0.31	0.35
γ	−0.52	0.12	0.78	−0.06	−0.24
δ_2	0.81	−0.36	0.09	0.37	−0.24
χ_2	0.83	−0.45	−0.01	0.29	0.02

PC#2 where Figure 4 indicates a larger contribution of ζ and Table III suggests a larger contribution of β. By taking the length of the loading vector into account one should have more confidence in the results of Figure 4. Although Table III suggests some contribution of β and γ to PC#1, their vector lengths are rather small as compared with the contributions by the χs and δs and ζ. Therefore, it can be concluded that β and γ do not contribute significantly to PC#1.

The cosine of the angle between the loading vectors should give an approximation of the correlation coefficient between the corresponding variables. Although the cosine of the angle can be calculated by means of eq. (6) we were first interested in the visual information resulting from the loading plot. In Table IV the expected size of the correlations and the sign between variables j and j', on the basis of the loading plots, are given. To get an impression of the suitability of torsion angle correlation estimation from loading vectors we introduced a scale indicating *low*, *middle*, and *high* values for the size of the correlation estimated from loading vectors.[‡] The corresponding calculated correlation coefficients, $r_{jj'}$, are also depicted.

The χs and δs are close together in the loading plots (i.e., a small positive angle). Hence, they have a high positive correlation, which is also expressed in their calculated correlation coefficients. Both the size and the sign of this correlation are correctly estimated from the plot. The estimation of both the size and the sign for the correlation and the corresponding calculated correlation coefficient for the other variable combinations compares well in most cases. Apart from the results in Table IV, the size and sign of the correlation between ζ and β are estimated to be ''high'' and ''+,'' respectively. The calculated correlation coefficient is **0.64**.

Hence, the results obtained in the previous paragraph can be specified further. The relation between the variables responsible for a certain DNA-class separation can now be expressed in terms of an estimated correlation coefficient. Not surprisingly, the high negative correlation between α and γ is again confirmed. The separation of B_I-DNA/B_{II}-DNA on PC#2 can now be specified. It involves the ϵ with ζ correlation, but obviously β is also highly correlated with both ϵ and ζ. In the analysis of the score plots, some correlation between χ_2 and PC#2 is found. However, Table IV shows no correlation between ϵ and χ_2 Hence, the correlation between χ_2 and PC#2 should be attributed to a χ_2−ζ correlation.

The separation of A-DNA from B-DNA clearly results from a difference in χs and δs but also from a difference in ζ. Besides the high positive correlation between χ and δ there is a negative correlation between the χs and δs with ζ. Analysis of the score plots also shows some correlation of β and γ with PC#1. Table IV shows no high correlation between χs or δs with either β or with γ.

Previous correlation studies on mononucleosides/nucleotides showed correlations between δ and χ but also between δ or χ with ϵ and between χ with ϵ. In the present study, δ and χ are clearly present and, to some extent, we see some correlation between χ and β. However we see no indication of a correlation between δ or χ with ϵ. The most general correlations found when studying polynucleosides/nucleotides are between α and β, α and γ, β and ϵ, and ϵ and ζ. Although the first one is detected in the present study it is not as prominent as the other three.

Biplots

Until now, SVD of the DM data matrix has led to a separation of DNA classes including an indication of the responsible variables (score plots), and an estimation of which variables are corre-

[‡] The table contains an arbitrary scale *high*, *middle*, and *low*. In this manner, the size of the correlation that is estimated from loading vectors can be classified. Because of this threefold division the scale corresponds to, respectively, $r_{jj'} > 0.66$, $0.33 \leq r_{jj'} \leq 0.66$, and $r_{jj'} < 0.33$. Strictly speaking, this means that a *high* correlation corresponds to an angle between vectors that is in the range of 0° to 49° or 131−180°, *middle* corresponds to an angle between 49° and 71° or 109° and 131°, and *low* corresponds to and angle between 71° and 109°.

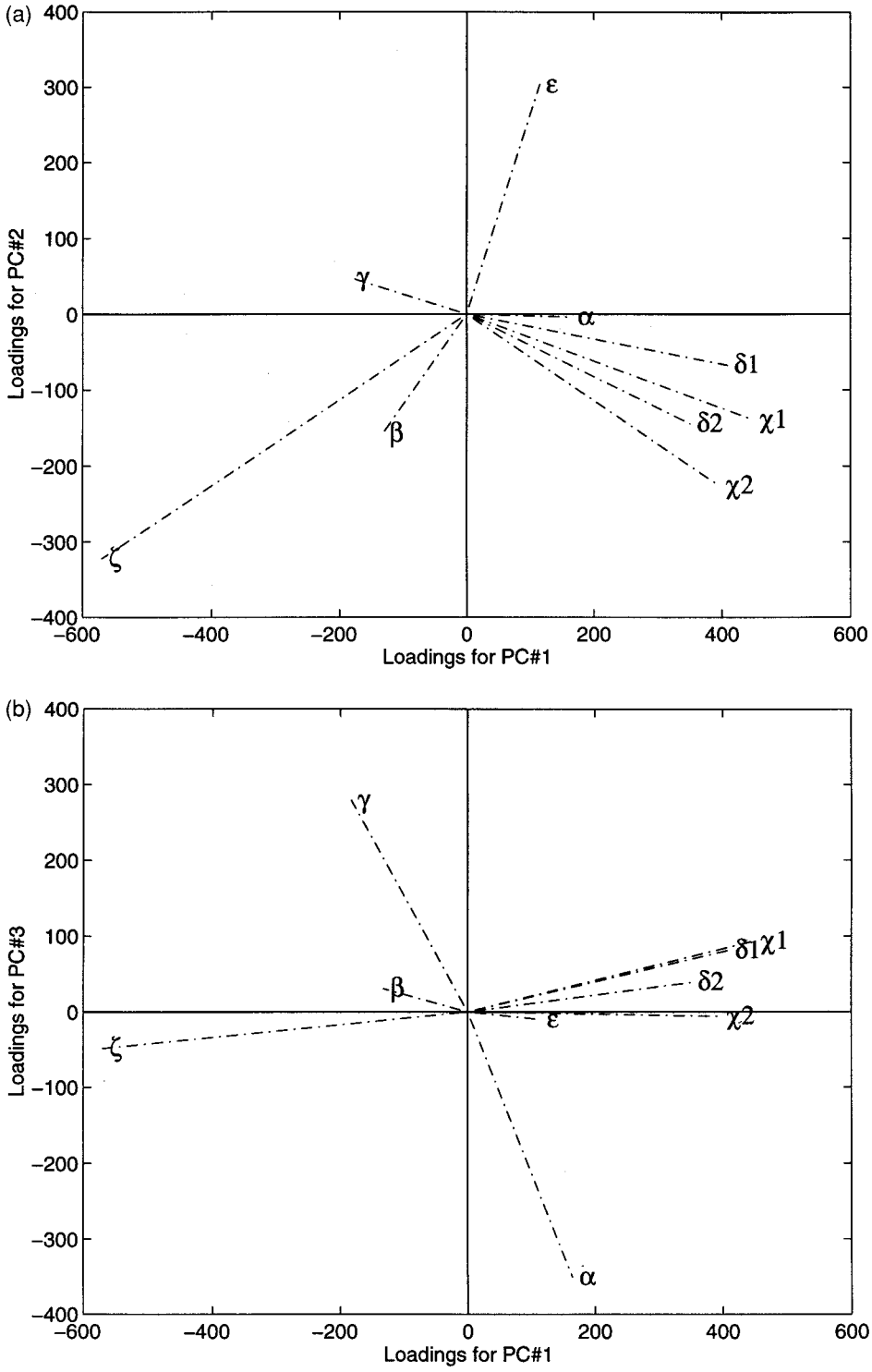

FIGURE 4. (a) Loadings of PC#1 against loadings of PC#2. (b) Loadings of PC#3 against loadings of PC#1.

TABLE IV.
Size and Sign of the Correlation between Variables Estimated from $\alpha = 0$ and $\beta = 1$ Loading Plots and Corresponding Calculated Correlation Coefficients.

	Variables	Size	Sign	Calculated $r_{jj'}$
PC#1	$\chi_1-\delta_1$	high	+	0.93
	$\chi_1-\epsilon$	low	−	0.02
	$\chi_1-\zeta$	high	−	−0.69
	$\chi_1-\alpha$	low	+	0.23
	$\chi_1-\beta$	low	−	−0.35
	$\chi_1-\gamma$	low	−	−0.35
	$\chi_1-\delta_2$	high	+	0.71
	$\chi_1-\chi_2$	high	+	0.77
PC#2	$\epsilon-\delta_1$	low	−	0.02
	$\epsilon-\chi_1$	low	+	0.08
	$\epsilon-\zeta$	high	−	−0.61
	$\epsilon-\alpha$	low	+	0.06
	$\epsilon-\beta$	high	−	−0.48
	$\epsilon-\gamma$	low	+	−0.14
	$\epsilon-\delta_2$	low	−	0.04
	$\epsilon-\chi_2$	middle	−	0.00
PC#3	$\alpha-\chi_1$	low	+	0.23
	$\alpha-\delta_1$	low	+	0.25
	$\alpha-\epsilon$	low	+	0.06
	$\alpha-\zeta$	low	−	−0.31
	$\alpha-\beta$	middle	−	−0.40
	$\alpha-\gamma$	high	−	−0.84
	$\alpha-\delta_2$	low	−	0.23
	$\alpha-\chi_2$	low	+	0.30

lated including the size of this correlation (loading plots). Instead of analyzing score and loading plots separately the same information should be obtained when using a single biplot. For this purpose the scores and loadings receive the same weight by applying eq. (7). As can be seen in Figure 5a, good class separation, as well as variable correlation estimation, can be achieved. An additional advantage of biplots is that one can see which DNA class(es) have high or low original variable values via the perpendicular projection of score points on the loading vectors. For example, consider the loading vector for ϵ in Figure 5a. Perpendicular projections from score points to this vector show that the B_{II}-DNA class has higher ϵ values than the A-DNA and B_I-DNA classes. Let us take the loading vector for χ_1. Imagine that this vector also continues in the opposite direction. Projecting scores on the vector shows that the A-DNA class has lower χ_1 values than the B-DNA classes. This can also be taken one step further and then one can see that the B_{II}-DNA class (on average) has even slightly higher χ_1 values than the B_I-DNA class. By proceeding in this way one obtains the data in Table V.

Apparently the A-DNA class has lower χ and δ and higher ζ values than the other classes. Moreover, a strong positive correlation between χ and δ and a strong negative correlation between ζ with both χ and δ is detected.

B_I-DNA has low ϵ and high ζ values, as opposed to B_{II}-DNA. Also, χ and δ appear to be somewhat lower for B_I-DNA than for B_{II}-DNA. Finally, β for B_I-DNA is higher than for B_{II}-DNA. The detected strong negative correlation between ϵ and ζ determines this class separation. As already seen, β has negative correlation with ϵ and positive correlation with ζ. Furthermore, the correlation of ζ with both χ and δ is detected.

The only variables responsible for the crankshaft effect are α and γ because the crankshaft objects only differ from other objects in that they have lower α values and higher γ values. Hence, biplot analysis strongly visualizes what is going on in DNA-class separation. To understand what is visualized in the biplots the results are explained in a more physical manner.

PHYSICAL INTERPRETATION

In an energetically favorable DNA-helix conformation there is hydrogen bonding between complementary bases, stacking[§] of succeeding bases in a single strand, and lack of van der Waals clashes. How can this be used to give a physical interpretation of the results that were found with the SVD procedure? Because single strand DM steps are used, the interpretation is only in terms of base stacking and van der Waals clashes.

In Figure 6, the $\chi-\delta$ a relation is depicted. By turning the δs in one direction the stacking of the bases is distorted (step 1). The stacking is re-established by turning the χs in the same direction (step 2). One can imagine that in step 2 some changes in the backbone might further improve base stacking. The biplot results indicate that ζ is

[§] Stacking of bases means that succeeding bases in a single strand (visually) are located parallel to each other. Although this would be energetically most favorable, there can, of course, be overlap due to numerous reasons.

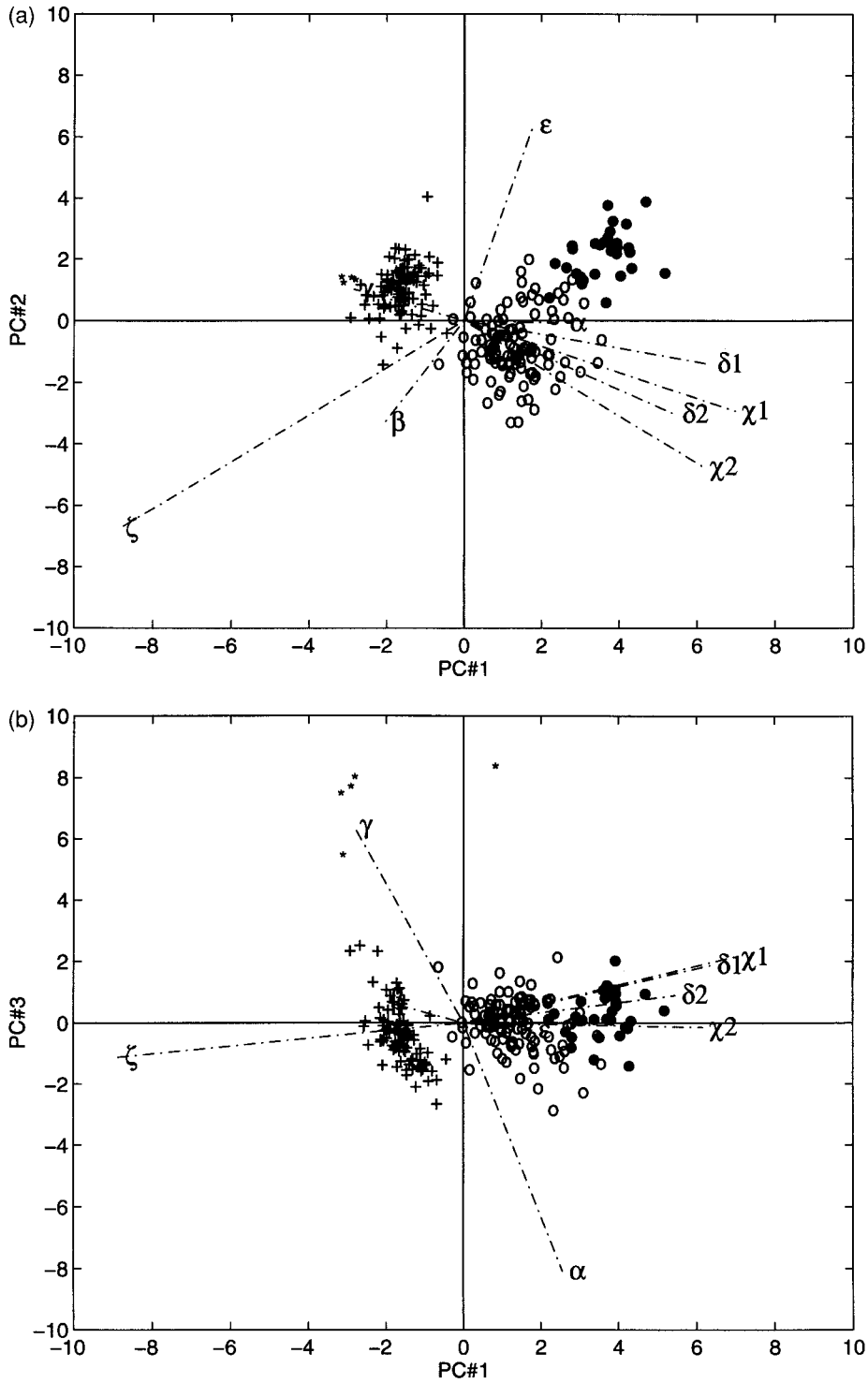

FIGURE 5. A-DNA (+); B_I-DNA (○); B_{II}-DNA (●); crankshafts (*). (a) SQ biplot of PC#1 against PC#2. (b) SQ biplot of PC#3 against PC#1.

TABLE V.
Intersections of Score Projections on Loading Vectors for Four DNA Classes.[a]

Variable	A-DNA	B_I-DNA	B_{II}-DNA	Crankshaft
χ_1	low	middle	high	low
δ_1	low	middle	high	low
ϵ	middle	low	high	middle
ζ	high	middle	low	high
α	middle	middle	middle	low
β	middle	middle	low	middle
γ	middle	middle	middle	high
δ_2	low	middle	high	low
χ_2	low	middle	high	low

[a] The indications ''high,'' ''middle,'' and ''low'' are relative for one loading vector, for example, the ζ values increase in the direction B_{II}-DNA, B_I-DNA, and A-DNA (average absolute values around 180°, 270°, and 290°, respectively). However, the α value only decreases for the crankshaft class to an (average) absolute value of about 180°, whereas the ''middle'' indication for all the other classes corresponds to an absolute value of about 300°.

the most probable torsion angle to aid in this improvement.

Figure 7 depicts the ϵ–ζ relation which is responsible for the B_I-DNA/B_{II}-DNA separation. When ζ moves from its normal $g-$ value toward a tr value, atom A approaches atom H2′ which inevitably results in a van der Waals clash (step 1). However, the molecule avoids this by, at the same time, moving ϵ in the opposite direction from its normal tr value toward a $g-$ value (step 2). It is known that this anticorrelated motion of ϵ not only avoids the van der Waals clash but also brings about changes in the helix axis. Hence, a compensation from other backbone torsion angles is needed to restore the helix. Biplots show that δ is involved here (therefore, in Fig. 7, δ, close to the ϵ–ζ couple, is also depicted). It was shown earlier in this study that a δ change also involves a change in χ—indeed, this is detected in the biplots. Finally, the biplots suggest an influence of β on the B_I-DNA/B_{II}-DNA separation. This is somewhat more difficult to interpret. A possible explanation might be that a crankshaft-like effect is involved (see hereafter); that is, to keep the α in a $g-$ region (as α does not change in a B_I-DNA/B_{II}-DNA transition), β might be decreased if ζ moves from $g-$ toward tr.

The effect, originally described as crankshaft, is depicted in Figure 8. If γ increases, α undergoes a corresponding decrease. Because the β torsion angle in between is usually tr, the energetically favorable parallelism of the bases is maintained. The biplots give no indication of other torsion angles involved in the crankshaft effect.[14, 23]

MULTIDIMENSIONAL RAMACHANDRAN PLOTS

Figure 2a and b depicts score plots of the DM data matrix after SVD. The formation of clusters of DMs indicates that certain areas in the r-dimensional factor space are more favorable than others. There may be areas that can never be accessed on the basis of physical grounds. The r-dimensional factor space after SVD is a subspace of the original n-dimensional torsion angles space. It seems of interest to transform the accessible and forbidden areas in the reduced space back to the original variables. In that case one could check whether torsion angle combinations would fall into an ac-

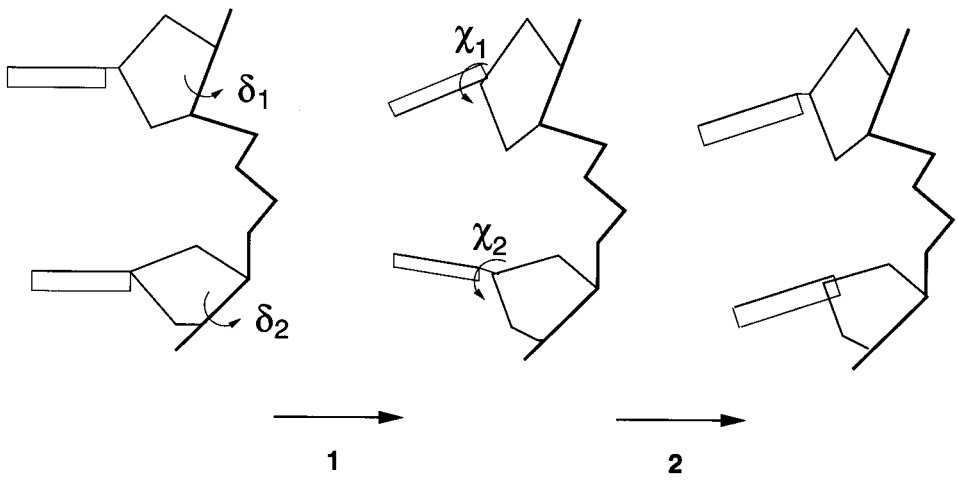

FIGURE 6. The χ is increased after a δ increase to re-establish base stacking (figure as in ref. 11).

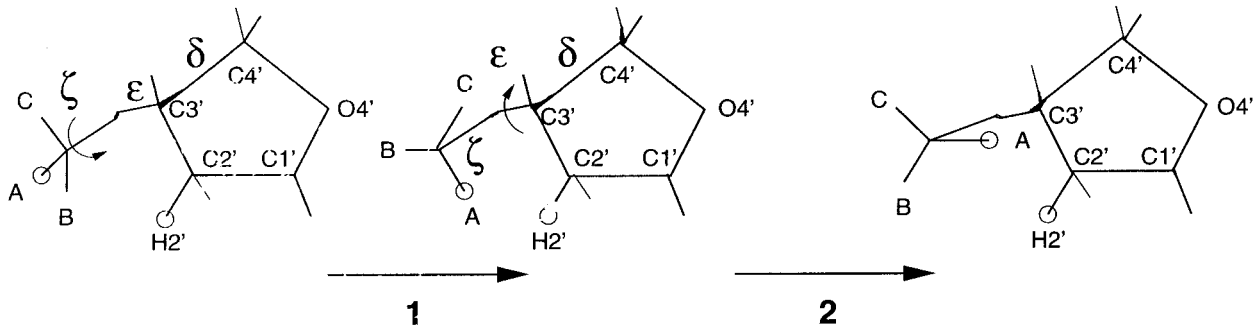

FIGURE 7. The ϵ is moved in the opposite direction from ζ in a B_I-DNA/B_{II}-DNA transition to avoid van der Waals clashes (figure as in ref. 11).

cessible or forbidden area. In such a way the equivalent of a Ramachandran plot (widely used in protein biology) for nucleic acids can be constructed.

The *REV* values given in Table II indicate that the number of independent variables in the data matrix is four; that is, a four-dimensional factor space results from the SVD. Hence, four PCs should suffice to represent the original data matrix and reconstruct the data matrix from the corresponding scores and loading matrices. However, visualization of a four-dimensional factor space is not possible. The first three PCs account for approximately 87% of the variance. The biplots indeed show that, with the first three PCs, all phenomena in the data matrix can be explained. Therefore, three PCs should give an acceptable reconstructed data matrix. Moreover, visualization of a three-dimensional factor space is possible; see Figure 9a ($\alpha = 0.5$).

According to $\hat{\mathbf{x}}_i = \mathbf{s}_i\mathbf{L}^T$, any chosen score vector, $\boldsymbol{s}_i$, can be reconstructed to an original variable vector, $\hat{\mathbf{x}}_i$. In this way one can select score vectors from a region between clusters in the score plot. Analysis of the reconstructed original variables may reveal why the specific area is considered forbidden. A few examples will be given.[‖] In Fig-

[‖] We also looked at the results of using λ as opposed to λ^2 in this calculation. The cumulative variance explained would be 32.8%, 50.8%, 75.0%, 85.5%,...,100%. Hence, although a smaller part of variance is captured with the first three PCs it is already sufficient to do a reasonable backcalculation of the torsion angles. Because we feel that the backcalculation on the basis of λ^2 is common we did not include the results on the basis of λ.

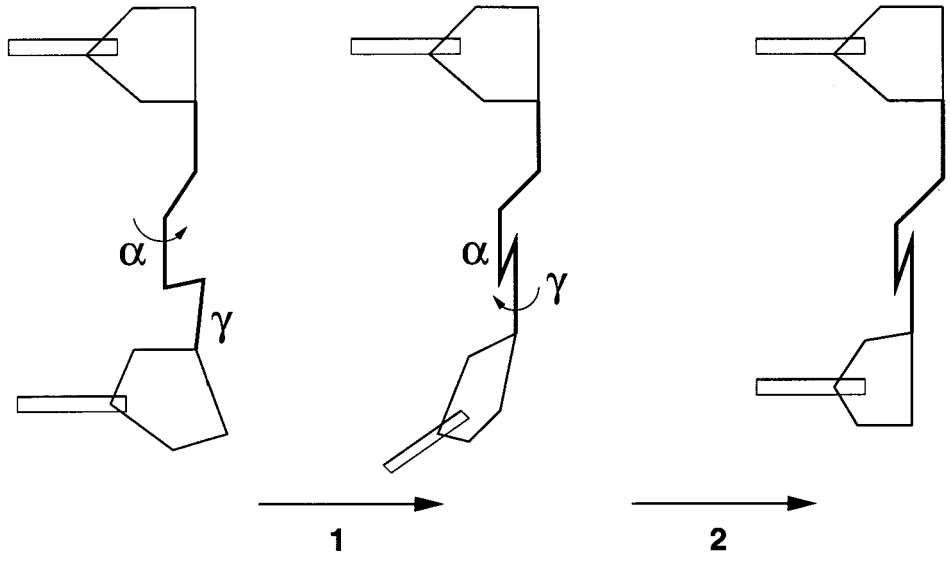

FIGURE 8. A crankshaft motion involves a decrease of α whenever γ increases to keep base stacking intact (figure as in ref. 11).

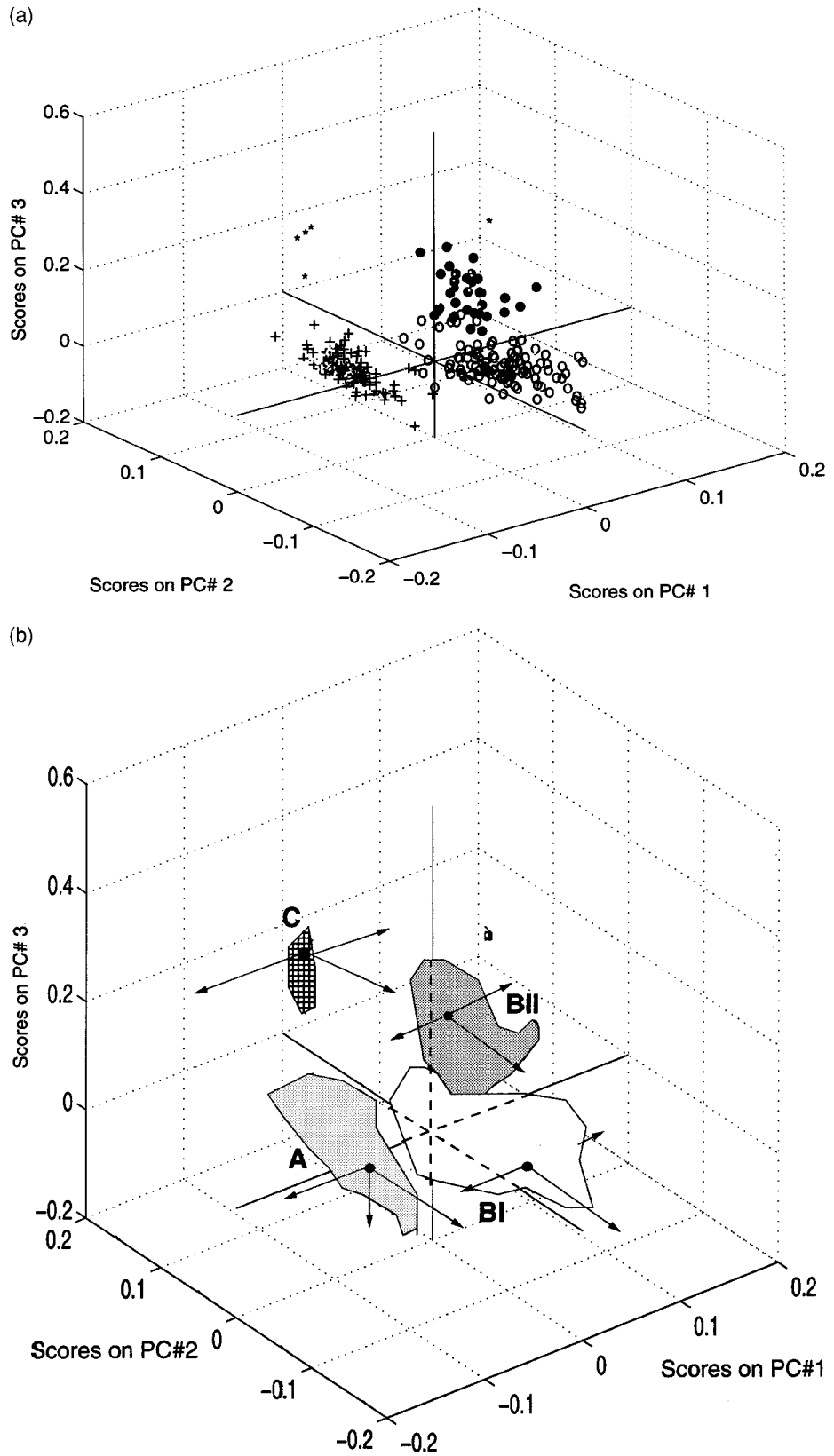

FIGURE 9. A-DNA (+); B_I-DNA (○); B_{II}-DNA (●); crankshafts (*). Scores on PC#1, PC#2, and PC#3. (b) A-DNA, B_I-DNA, B_{II}-DNA, and crankshaft classes are depicted by the shaded areas in the three-dimensional score plot of the first three PCs.

ure 9b, the DNA classes are indicated by shaded areas. In each class a representative DM score is depicted by a black dot.

First, the representative DM from the B_I class is considered. The score coordinates for this DM are the first entry in Table VI. The reconstructed torsion angles show a commonly found B_I-DNA conformation. What happens if new coordinates are chosen on the right of the B_I-DNA cluster (PC#1 value becomes 0.15)? The reconstructed torsion angles have rather high χ and δ values, but the most striking aspect is the combination of ϵ with ζ. ζ is decreased toward a *tr* value, but the corresponding increase of ϵ to $g-$ did not take place. Hence, according to Figure 7, a van der Waals clash is the result. Certainly, this clash cannot be avoided by the observed increase of χ and δ. If coordinates are chosen on the left of the B_I-DNA cluster (PC#1 value becomes -0.15) the reconstructed ζ has a value higher than 360. Therefore, ζ may be considered to have moved toward a low value in the $g+$ region. The expected increase in ϵ does not take place. Hence, a clash can also be expected here. For coordinates in front of the B_I-DNA cluster (PC#2 value becomes -0.20) a very low value of ϵ is reconstructed. In ref. 5 it is demonstrated that this will result in considerable steric hindrance whatever the value of other torsion angles may be.

In the same manner we can proceed for scores in and around the other clusters. New coordinates on the right of the B_{II}-DNA cluster still lead to reconstructed torsion angles with a B_{II}-DNA-like conformation. New coordinates on the left of the B_{II}-DNA cluster result in a less pronounced B_{II} $\epsilon-\zeta$ combination together with a A-DNA-like χ and δ. Because this contradicts the negative $\zeta-\delta$ correlation it will probably lead to steric hindrance. For coordinates in front of the B_{II}-DNA cluster the B_I coordinates are approached and indeed reconstructed B_I torsion angles are found.

Although new coordinates on the left of the A-DNA cluster depict the expected negative $\zeta-\delta$ a correlation, the reconstructed ζ is rather high. If coordinates are chosen below the cluster the resulting torsion angles still have a clear A-DNA conformation. The strong negative $\alpha-\gamma$ correlation is shown here. The presence of new coordinates in front of the A-DNA cluster results in an unacceptable $\epsilon-\zeta$ combination.

Reconstructed torsion angles for coordinates on the left of the crankshaft DMs show a more pronounced A-DNA conformation. For coordinates on the right side a B_{II}-DNA conformation is the result. The negative $\zeta-\delta$ correlation is in effect here. As was shown earlier in this study there is one B-DNA DM that has a crankshaft $\alpha-\gamma$ combination. Probably, the acceptable area for B-DNA hav-

TABLE VI.
Reconstructed Torsion Angle Values by Making Use of First Three PCs.

DNA type	PC#1	PC#2	PC#3	χ_1	δ_1	ϵ	ζ	α	β	γ	δ_2	χ_2
B_I	[0.02	−0.10	0.02]	261	126	171	292	290	188	51	130	266
	[0.15	−0.10	0.02]	320	179	186	218	311	171	27	175	318
	[−0.15	−0.10	0.02]	185	56	151	389	262	210	82	70	199
	[0.02	−0.20	0.02]	275	132	140	324	290	204	46	144	289
B_{II}	[0.08	0.12	0.06]	261	138	244	185	285	146	61	121	240
	[0.15	0.12	0.06]	293	167	252	145	296	137	49	145	268
	[−0.05	0.12	0.06]	202	85	229	258	263	164	85	75	189
	[0.08	−0.15	0.06]	295	155	171	262	286	184	50	155	294
A	[−0.06	0.02	−0.02]	205	82	198	301	290	178	60	83	208
	[−0.15	0.02	−0.02]	164	45	188	352	275	190	76	51	172
	[−0.06	0.02	−0.15]	193	71	200	307	336	174	24	78	209
	[−0.06	−0.15	−0.02]	223	93	147	356	291	205	52	107	246
C[a]	[−0.10	0.05	0.35]	217	93	199	296	154	190	172	79	183
	[−0.15	0.05	0.35]	195	73	193	324	145	196	181	61	163
	[0.15	0.05	0.35]	330	195	228	151	194	157	126	166	282
	[−0.10	0.00	0.35]	224	96	184	312	154	198	170	86	194

[a] Crankshaft.

ing a crankshaft effect is in the B_{II} domain. Finally, when taking coordinates in front of the crankshaft cluster an A-DNA conformation occurs.

Obviously, points in score space that are located near class borders (or intermediate class borders) are most sensitive for errors in torsion angle reconstruction. However, because, by means of three PCs, almost 90% of the variance is captured, for most of the score points there will be no serious problems in this kind of reconstruction.

Conclusions and Outlook

We applied singular value decomposition (SVD) to a data matrix containing 244 DMs represented by nine torsion angles. As opposed to calculating pairwise correlation coefficients, which is often presented in the literature, this multivariate approach immediately reveals multiple relations. Moreover, score plots resulting from SVD provide the basis for an equivalent of a multidimensional Ramachandran plot for nucleic acids. In contrast to the original Ramachandran plot for proteins that considers only two torsion angles, the Ramachandran plot derived from the SVD score plot, depicted in three dimensions in this study, indirectly concerns all original n (> 2) torsion angles.

Especially when using biplots an analysis of how torsion angles are related to each other and with class separation is easily visualized. It is known that A-DNA is characterized by lower χ and δ values than for B-DNA. This class separation is shown by the SVD results. However, an additional feature that can be detected in the biplots is the way ζ is involved in this relation. B_I-DNA and B_{II}-DNA are distinguished from each other by their ϵ and ζ values. Besides this correlation, biplots show that β is also involved and that the $\chi-\delta$ relation plays a role. Biplots indicate no additional torsion angles, other than the known $\alpha-\gamma$ relation, which are responsible for the separation of a crankshaft class. The class separation and the way torsion angles are related to this separation are also illustrated physically in this study.

Analyzing a correlation matrix of pairwise correlation coefficients between torsion angles certainly reveals which (groups of) torsion angles are related. However, it is difficult to interpret these relations physically on the basis of such a matrix alone. By using the results of a multivariate approach (e.g., SVD), and exploiting the strong visualizing abilities of biplots, one is able to directly relate the results to a physical interpretation. Moreover, scores resulting from SVD can be reconstructed directly into torsion angle values. This means that a distinction between accessible and forbidden areas in a low dimensional score plot can be translated into the corresponding areas in the original full-dimensional torsion angle space. Hence, the score plots can be used as a multidimensional equivalent of a Ramachandran plot.

Although the principle is clearly outlined in this article, a more efficient use of such Ramachandran-like plots probably requires more DMs and a larger diversity of DMs. More DMs provide the basis for elaborate cluster/classification algorithms that are capable of estimating the volume of clusters. This may lead to better defined accessible and forbidden areas. A larger diversity of DMs may lead to additional clusters (such as sharp hairpin turns, etc.). One might also look at other formulations of SVD to improve the class separation even further.

Acknowledgments

The authors thank Drs. S. Pontfoort for preliminary investigations on the subject. Dr. E. P. P. A. Derks and Dr. R. Wehrens are acknowledged for fruitful discussions. Dr. M. M. W. Mooren is acknowledged for providing the figures and information depicted in the Appendix.

Appendix

Figure A-1 a schematic representation of a flexible chain defining different orders of steric contacts generated by different combinations of torsion angles such as those used in ref. 5. A variable torsion angle, θ, in the nucleoside/nucleotide chain is indicated by an arrow, whereas a fixed torsion angle is indicated by an ''F.'' Filled circles represent atoms that are involved in steric interactions when the values of the variable torsion angles are changed.

For each atom pair possibly involved in steric interaction, within the framework of a particular order of interaction, a separate contour map has to be calculated. Steric interactions occur when distances between the atoms (''black atoms'' in Fig. A-1) become shorter than the chosen, permitted van der Waals contact distances. These contact distances are, in this case, chosen to be 0.1–0.2 Å shorter than the ''normally allowed'' distances (Table A-1).

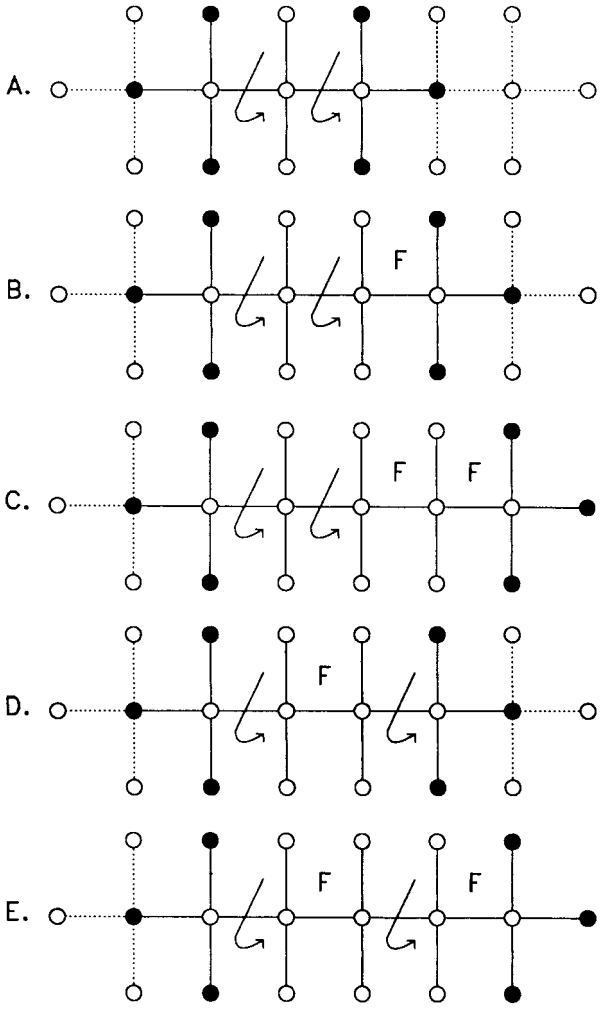

FIGURE A-1. (A) Representation of second order steric contacts $\theta - (\theta + 1)$. (B) Representation of third order steric contacts $\theta - (\theta + 1) - [(\theta + 2)]$. (C) Representation of fourth order steric contacts $\theta - (\theta + 1) - [(\theta + 2)] - [(\theta + 3)]$. (D) Representation of third order steric contacts $\theta - [(\theta + 1)] - [(\theta + 2)]$. (E) Representation of fourth order steric contacts $\theta - [(\theta + 1)] - (\theta + 2) - [(\theta + 3)]$. (From Mooren, Thesis, University of Nijmegen, Nijmegen, The Netherlands, © 1993, reproduced with permission.)

Torsion angles were varied between 0° and 360° in steps of 18° while bond lengths and bond angles were kept fixed at values provided in ref. 23. As an example, the conformational space allowed for the torsion angles α_i and β_i is presented in Figure A-2. When only one second order contact OA_i—$C4'_i$ is taken into account, one derives the plot in the left panel. The contours represent lines between calculated points for which the distance from OA_i—$C4'_i$ is constant. The shaded area indicates for which torsion angle combination the distance be-

TABLE A-I.
Atom Pairs Possibly Involved in Steric Interaction and the Corresponding Contact Distance.

Atom pair	Contact distance (Å)
C—C	3.0
C—O	2.7
C—N	2.8
C—P	3.2
C—H	2.2
O—O	2.7
O—N	2.6
O—P	3.0
O—H	2.2
N—N	2.6
N—P	3.0
N—H	2.2
P—P	3.3
P—H	2.5
H—H	1.9

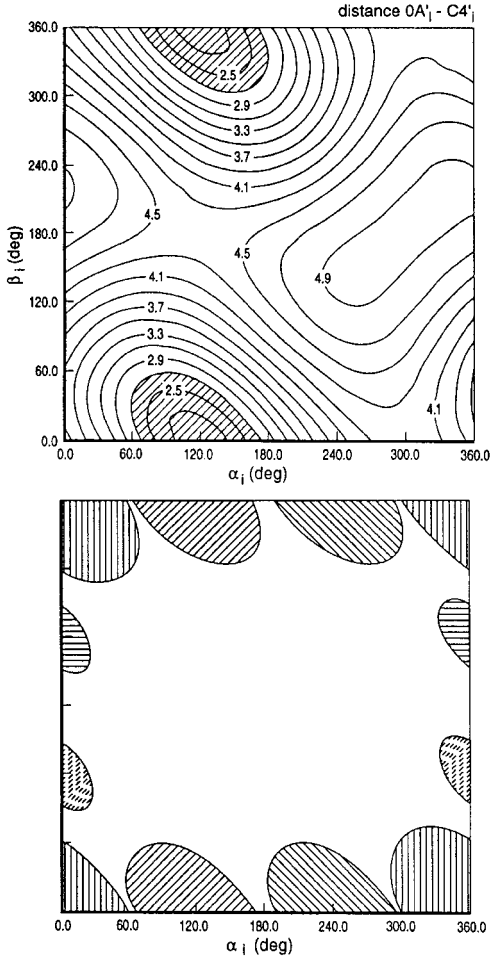

FIGURE A-2. Contour lines representing constant OA_i–$C4'_i$ distances for $\alpha_i - \beta_i$ torsion angle combination (left top panel). Contour plot with prohibited (shaded areas) and allowed areas in $\alpha_i - \beta_i$ combination space (bottom). (From Mooren, Thesis, University of Nijmegen, Nijmegen, The Netherlands, © 1993, reproduced with permission.)

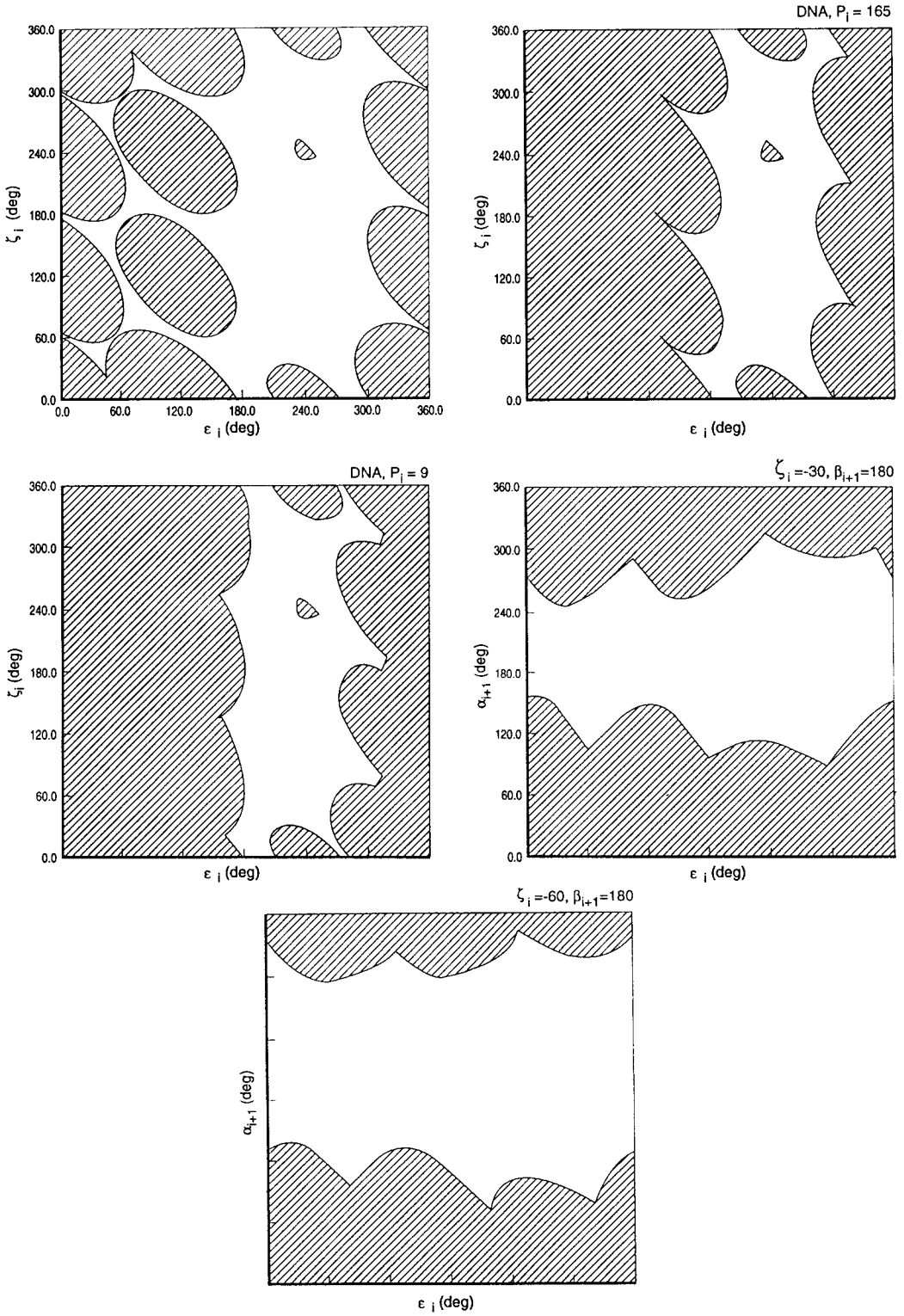

FIGURE A-3. Contour plots for several combinations of torsion angles. (From Mooren, Thesis, University of Nijmegen, Nijmegen, The Netherlands, © 1993, reproduced with permission.)

tween the oxygen and the carbon becomes less than the allowed 2.7 Å. When the same procedure is followed for the remaining atom pairs one can overlay all maps and derive the figure with prohibited zones depicted in the right panel (Fig. A-2). Figure A-3 shows examples of such maps.

References

1. R. E. Dickerson et al., *J. Mol. Biol.*, **205**, 787 (1989).
2. A. A. Gorin, V. B. Zhurkin, and W. K. Olson, *J. Mol. Biol.*, **247**, 34 (1995).
3. M. A. El Hassan and C. R. Calladine, *J. Mol. Biol.* **259**, 95 (1996).
4. G. N. Ramachandran, C. Ramakrishan, and V. Sasisekharan, *J. Mol. Biol.*, **7**, 95 (1963).
5. M. W. M. Mooren, *On Nucleic Acid Structure Analysis by NMR*, Thesis, University of Nijmegen, Nijmegen, The Netherlands, 1993.
6. M. Sundaralingam, *Biopolymers*, **7**, 821 (1969).
7. W. K. Olson and P. J. Flory, *Biopolymers*, **11**, 1 (1972).
8. W. K. Olson, *Biopolymers*, **15**, 859 (1976).
9. N. Yathindra and M. Sundaralingam, *Nucl. Acids Res.*, **3**, 729 (1976).
10. C. Altona and M. Sundaralingam, *J. Am. Chem. Soc.*, **94**, 8205 (1972).
11. A. V. Fratini, M. L. Kopka, H. R. Drew, and R. E. Dickerson, *J. Mol. Chem.*, **247**, 14686 (1982).
12. B. N. Conner, C. Yoon, J. L. Dickerson, and R. E. Dickerson, *J. Mol. Biol.*, **174**, 663 (1984).
13. B. Schneider, S. Neidle, and H. M. Berman, *Biopolymers*, **42**, 113 (1997).
14. D. A. Pearlman and S.-H. Kim, *J. Biomol. Struct. Dynam.*, **4**, 49 (1986).
15. G. G. Privé, U. Heinemann, S. Chandrasegaran, L. S. Kan, M. L. Kopja, and R. E. Dickerson, *Science*, **238**, 498 (1987).
16. H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. Hsiek, A. R. Srinivasan, and B. Schneider, *Biophys. J.*, **63**, 751 (1992). Internet: http://ndb-server.rutgers.edu/
17. E. J. Jackson, *A User's Guide to Principal Component Analysis*, John Wiley & Sons, New York, 1990.
18. K. R. Gabriel and C. L. Odoroff, *The Biplot for Exploration and Diagnosis: Examples and Software*, Department of Statistical Reporting #86/03, University of Rochester, Rochester, NY, 1986.
19. P. J. Lewi, *Arzneim. Forsch.* (*Drug. Res.*), **26**, 1295 (1976).
20. P. J. Lewi, *Chem. Intell. Lab. Syst.*, **5**, 105 (1989).
21. R. D. Catell, *Multivariate Behavioral Research*, **1**, 245 (1966).
22. E. R. Malinowski, *J. Chemometrics*, **1**, 33 (1987).
23. W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984.